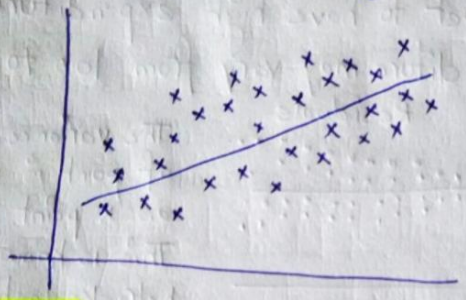
REGRESSION ASSUMPTIONS

All assumptions should be followed before model:

i) Linearity ii) Constant Error variance iii) Independent error terms iv) Normal errors
v) No multi collinearity vi) Exogeneity



$\hat{y} = 3.1 + 0.79x$.

	Coef	Std Error	t Stat	P-value
Intercept	3.605	1.4024	2.1823	0.2909
x	0.7393	0.2308	3.2033	0.0013

i) Linearity

- Regression should be linear in terms of β.
- It means it can have an additive regression equation so it variables can be added easily like $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \beta_n x_n + \varepsilon$.

ii) Constant Error variance (Homoscedacity) / (No heteroskedasticity)

- Variance in the errors (distance between dot and line) try to remain constant throughout the line.

iii) Independence error terms (Auto correlation)

- When each successive error is independent of last one.

iv) Normal errors

- Spread of error should be normally distributed (bell shaped).

v) No multicollinearity (Truly independent x terms)

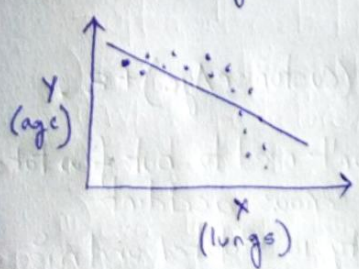
- Multicollinearity occurs when X variables are themselves related.

vi) Exogenity

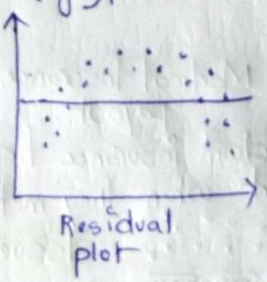
- Omitted variable bias.

1. LINEARITY (Correct functional form)

Consider Lung function = $\beta_0 + \beta_1 (age)_i + \varepsilon_i$



Residual plot → (make line parallel)



Lung function = $\beta_0 + \beta_1 (age)_i + \beta_2 (age^2)_i$

Residual plot

Issue → If the functional form is incorrect, both the coefficients and standard error in your output are unreliable.

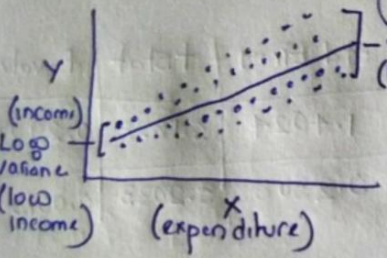
Detection of Linearity – i) Residual plots
ii) Likelihood ratio (LR) test.

2) Constant Variance (homoscedacitity / no heteroskedasticity)

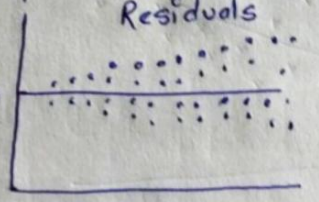
Consider the following model, Expenditure = $\beta_0 + \beta_1$ (income) + ε_i.

When income is low, we do not have budget to have high expenditure.

But when income is high, budget of expenditure may vary from low to high.



Residual plot →



The variance or spread of point is increasing as x point increasing. This is known as heteroskedasticity.

Hetro means numerous, skedasticity meaning variance

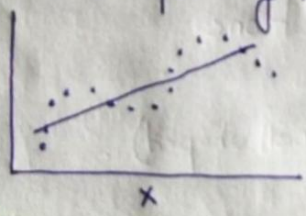
Issue → Under heteroskedasticity, standard error in output cannot be relied upon.

Detection → Goldfeldt-Quant test.
Breusch-Pagan test

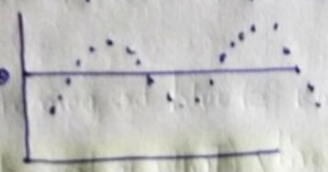
Remedies – i) White's standard error
ii) Weighted least squares.
iii) Log things (log variable, transform)

3) Independence error terms (No auto correlation)

Consider the following model, stock Market Index = $\beta_0 + \beta_1$ (Time) + ε_i



Residual plot →



Issue → Under autocorrelation, standard errors in output cannot be relied upon.

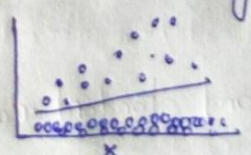
Detection → Durbin – Watson test
Breusch – Godfrey test

Remedies – i) Investigate omitted variables.
ii) Generalised difference equation (Cochrane – Orchutt / ARIMA methods)

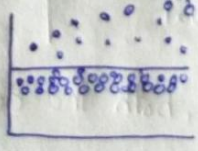
4) Normality of errors

Consider the following model, Medical Insurance payout = $\beta_0 + \beta_1$ (Customer Age) + ε.

Normally most of the people don't claim insurance because they don't need to but few take out and take huge sum of amount when they met with some serious accident



Residual plot →



Issue – If normality is violated and n is small, standard errors in output are affected. (Komolgorov-Smirnov test)

Detection – i) Histogram / QQ plot
ii) Shapiro-Wilk test
iii) KS test
iv) Anderson-Darlight

Remedies – i) Change functional form (log)

5) No Multicollinearity

Consider the following models, Motor Accident $= \beta_0 + \beta_1$ (Num cars) $+ \beta_2$ (No of residents) $+ \varepsilon$

Multi-collinearity occurs when the X variables are themselves related.

Issue → i) Coefficients and standard errors of affected variables are unreliable.

Detection → i) Look at correlation (ρ) between X variables.
ii) Look at Variance Inflation factor (VIF)

Remedies → Remove one of the variable.

NOTE - Adding an interaction term will not fix the problem. Suppose we make a new variable which is combination of cars and resident (suppose multiplication) that will not solve a problem.

6) Exogenity (no omitted variable bias)

Consider the following model, Salary $= \beta_0 + \beta_1$ (Years of education) $+ \varepsilon_i$

Here there are other variable also which can affect salary like socio-economic status like family income which can afford the cost of study which lead to many times quality of education.

So, socio-economic status affects both X and Y variables, thus could cause omitted variable bias.

Technically, socio-economic status would effect ε_i in the model thus education is no longer wholly exogenous as it can be explained in part by the error term.

Issue → Model can only used for predictive purpose (cannot infer causation)

Detection → i) Intution ii) Checking correlation

Remedy → Using instrumental variables.